Psihologijske teme, 34 (2025), 3, 581-606

Pregledni rad UDK: 159.9.072

https://doi.org/10.31820/pt.34.3.10

Što čestice doista mjere? Uvod u teoriju odgovora na zadatak

Ana Ćosić Pilepić

Sveučilište u Rijeci, Filozofski fakultet, Odsjek za psihologiju, Rijeka, Hrvatska

Sažetak

Teorija odgovora na zadatak (IRT) suvremeni je psihometrijski okvir koji modelira vjerojatnost odgovora na pojedine čestice kao funkciju jedne ili više latentnih osobina. Ovaj rad donosi pregled teorijskih temelja, pretpostavki i primjena IRT-a, s posebnim naglaskom na njegovu upotrebu u psihološkome mjerenju. Cilj je rada prikazati ključne IRT modele za dihotomne i politomne čestice, usporediti ih s klasičnom teorijom testova te raspraviti o njihovim prednostima. U radu se objašnjavaju osnovni parametri čestica (težina, diskriminativnost, pogađanje), opisuju se karakteristične krivulje čestica i testa te analiziraju temeljne pretpostavke poput jednodimenzionalnosti i lokalne nezavisnosti. Poseban je naglasak stavljen na primjenu IRT-a u različitim područjima psihologije te na njezine doprinose psihometrijskoj praksi. Osim invarijantnosti parametara, preciznosti mjerenja i mogućnosti detekcije diferencijalnoga funkcioniranja čestica kao prednosti IRT-a, razmatraju se i izazovi primjene IRT-a, poput potrebe za većim uzorcima, složenosti analitičkih postupaka i ograničenja u kontekstu mjerenja tipičnoga ponašanja. Zaključno, rad pokazuje da je IRT, unatoč praktičnim zahtjevima, temeljni alat za razvoj suvremenih psiholoških mjernih instrumenata jer osigurava veću točnost i znanstvenu utemeljenost psihologijskoga mjerenja.

Ključne riječi: teorija odgovora na zadatak, psihometrija, modeli za dihotomne čestice, modeli za politomne čestice, psihološko mjerenje

Uvod

Psihološko mjerenje neizostavan je dio teorijskoga i primijenjenog rada u psihologiji. Bilo da je riječ o procjeni osobina ličnosti, sposobnosti ili kliničkih simptoma, kvaliteta psihometrijskih instrumenata izravno utječe na valjanost zaključaka koje istraživači i praktičari donose. Tijekom većega dijela 20. stoljeća klasična teorija testova (engl. *Classical Test Theory*, CTT) bila je dominantni okvir za konstruiranje, evaluaciju i primjenu testova. Međutim, njezina konceptualna i statistička ograničenja, posebno ovisnost metrijskih karakteristika o uzorku i nemogućnost detaljne analize na razini pojedinačnih čestica, potaknuli su razvoj modernih psihometrijskih pristupa.

Ana Ćosić Pilepić 0000-0001-6780-8600

✉ Ana Ćosić Pilepić, Sveučilište u Rijeci, Filozofski fakultet, Odsjek za psihologiju, Sveučilišna avenija 4, 51 000 Rijeka, Hrvatska. E-adresa: *acosic@uniri.hr*

Jedan je od najvažnijih suvremenih pristupa psihometriji teorija odgovora na zadatak (engl. *Item Response Theory*, IRT) koja donosi fundamentalno drugačiji pogled na odnos između latentne osobine i odgovora na čestice. Za razliku od CTT-a koji se temelji na ukupnome rezultatu, IRT modelira odgovore na pojedinačne čestice pomoću probabilističkih funkcija, uzimajući u obzir parametre karakteristične za čestice i latentne osobine ispitanika (Reckase, 2009).

Osnovna je pretpostavka IRT-a da se odgovor na svaku česticu može predvidjeti na temelju jednoga latentnoga konstrukta (ili više njih), a funkcija vjerojatnosti odgovora određena je parametrima težine, diskriminativnosti i ponekad pseudopogađanja te parametrom gornje asimptote. Takav pristup omogućuje preciznije mjerenje, analizu informativnosti čestica te procjenu pouzdanosti koja varira ovisno o razini latentne osobine – što klasičnim metodama nije moguće (Embretson i Reise, 2000).

Cilj je ovoga preglednog rada pružiti sustavan prikaz teorije odgovora na zadatak kao suvremenoga psihometrijskog pristupa, s naglaskom na njezine temeljne koncepte i modele, potom prednosti i nedostatke u odnosu na klasičnu teoriju testova te primjere primjene u psihologiji. U radu se pojašnjavaju osnovni parametri čestica (težina, diskriminativnost, parametar pogađanja i gornje asimptote), opisuju se koncept informacijske funkcije čestice i testa te standardna pogreška mjerenja, a detaljnije se razrađuju najvažniji modeli za dihotomne i politomne čestice. Budući da su dihotomni modeli polazište za razumijevanje modela za politomne čestice, prvi su detaljnije prikazani, dok se za politomne modele donosi osnovni pregled. Posebna je pažnja posvećena usporedbi s CTT-om, uključujući teorijske i praktične prednosti IRT-a, ali i potencijalna ograničenja u primjeni na empirijske podatke. Rad je ponajprije namijenjen psiholozima, istraživačima i praktičarima koji žele bolje razumjeti mogućnosti koje IRT pruža u konstrukciji, primjeni i evaluaciji psiholoških mjernih instrumenata.

Osnovne pretpostavke i koncepti

Teorija odgovora na zadatak usmjerena je na modeliranje odnosa između odgovora ispitanika na pojedinoj čestici i latentne osobine (konstrukta) koju taj instrument mjeri. Ta latentna osobina, simbolički označena s θ (grč. theta), predstavlja sposobnost, znanje, stav ili osobinu ispitanika (Crocker i Algina, 2008).

IRT pretpostavlja da su odgovori ispitanika ponajprije određeni njihovom razinom latentne osobine (npr. emocionalna inteligencija) te karakteristikama same čestice. Primjerice, osoba s višom razinom emocionalne inteligencije imat će veću vjerojatnost da na tvrdnju poput „Dobro raspoloženje mogu zadržati čak i kada mi se dogodi nešto loše" (Takšić, 2002) odabere odgovor „5 – *u potpunosti DA*". S druge strane, osoba s nižom razinom emocionalne inteligencije ima veću vjerojatnost da će odabrati niži stupanj slaganja, čime iskazuje nižu razinu sposobnosti koju skala mjeri.

Koje će karakteristike čestica biti uključene u modeliranje odgovora ispitanika,

ovisi o razini složenosti odabranoga modela, a izbor te složenosti određuje se prema obilježjima samoga mjernog instrumenta. Tako se čestice mogu modelirati s jednim parametrom ili više njih, najčešće onima koji označavaju težinu i diskriminativnost čestice, što su pokazatelji važni i u CTT-u. U složenijim modelima uključuju se i dodatni parametri, poput vjerojatnosti točnoga odgovora pri vrlo niskim razinama latentne osobine (tzv. pseudopogađanje) ili maksimalne vjerojatnosti točnoga odgovora (tzv. *slipping* parametar ili parametar gornje asimptote). Upravo uvođenje tih parametara omogućuje jednu od ključnih prednosti IRT-a u odnosu na CTT: mogućnost grafičkoga prikaza svojstava svake čestice.

IRT, osim toga, omogućuje da se i ispitanici i čestice smjeste na istoj mjernoj ljestvici latentne osobine, tj. i težina čestice i razina osobine koju pojedinac posjeduje izražavaju se u istim jedinicama. Nasuprot tomu, u okviru klasične teorije testova ne postoji mogućnost izravne usporedbe težine čestice i razine latentne osobine.

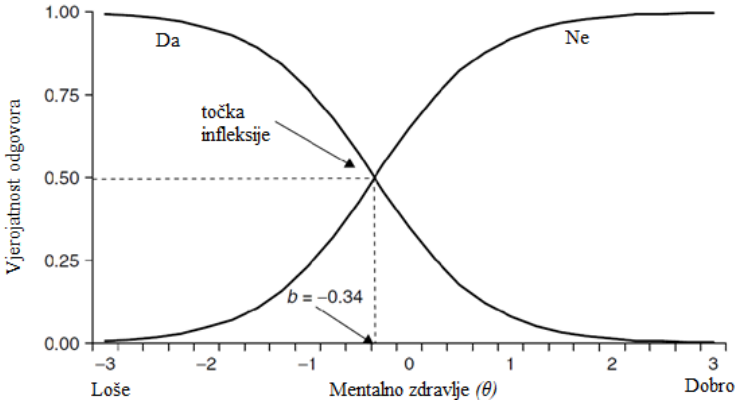
U nastavku su prikazani temeljni koncepti koje je važno razumjeti za primjenu IRT-a, počevši od karakteristične krivulje čestice, središnjega elementa većine IRT modela.

Karakteristična krivulja čestice

Povezanost između odgovora ispitanika na pojedinu česticu i njegove latentne osobine modelira se putem tzv. karakteristične krivulje čestice (engl. *item characteristic curve*, ICC). Ta je krivulja poznata još i kao linija traga čestice (engl. *item trace line*), krivulja funkcije odgovora na česticu (engl. *item response function*, IRF) ili krivulja kategorije odgovora (engl. *category response curve*, CRC) u slučaju politomnih čestica (deMars, 2010).

Slika 1.

Karakteristične krivulje odgovora na česticu „Jeste li u protekla 4 tjedna zbog utjecaja bilo kakvih emocionalnih problema (npr. osjećaj depresije ili tjeskobe) na poslu ili pri obavljanju nekih drugih svakodnevnih aktivnosti obavili manje nego što ste željeli?"



Napomena. Težina čestice od -0.34 predstavlja razinu mentalnoga zdravlja pri kojoj je vjerojatnost odgovora „da" i „ne" jednaka (tj. 50 %). Prilagođeno prema Chang i Reeve, 2005.

Na Slici 1. prikazan je primjer karakteristične krivulje čestice preuzet iz supskale psihičkoga zdravlja Upitnika zdravstvenoga statusa SF-36 (Ware i Sherbourne, 1992). Konkretna čestica glasi: „Jeste li u protekla 4 tjedna zbog utjecaja bilo kakvih emocionalnih problema (npr. osjećaj depresije ili tjeskobe) na poslu ili pri obavljanju nekih drugih svakodnevnih aktivnosti obavili manje nego što ste željeli?". Latentna je varijabla koja se mjeri tom supskalom psihičko zdravlje (θ), prikazano na horizontalnoj osi grafikona. Osobe s lošijim samoprocijenjenim psihičkim zdravljem nalaze se na lijevoj strani kontinuuma, dok se one s boljim psihičkim zdravljem pozicioniraju desno.

Vrijednosti na apscisi predstavljaju standardizirane jedinice, dok ordinata prikazuje vjerojatnost (0 – 1) da će osoba odabrati određenu kategoriju odgovora na čestici. U ovome su primjeru prikazane dvije karakteristične krivulje koje modeliraju vjerojatnost odgovora „da" i „ne", ovisno o razini latentne osobine. Očekivano, vjerojatnost odgovora „ne" raste kod osoba s višom razinom psihičkoga zdravlja, dok je vjerojatnost odgovora „da" viša kod onih s lošijim samoprocijenjenim zdravljem.

Zbog komplementarnosti funkcija kategorija kod dihotomnih čestica u praksi se često eksplicitno prikazuje samo jedna (najčešće rastuća) funkcija. Takve krivulje omogućuju vizualno i kvantitativno razumijevanje načina na koji čestica „funkcionira", pokazujući pri kojim razinama latentne osobine najviše razlikuje ispitanike i pružajući najviše informacija. Detaljniji matematički prikaz tih funkcija i modela donosi se u nastavku rada.

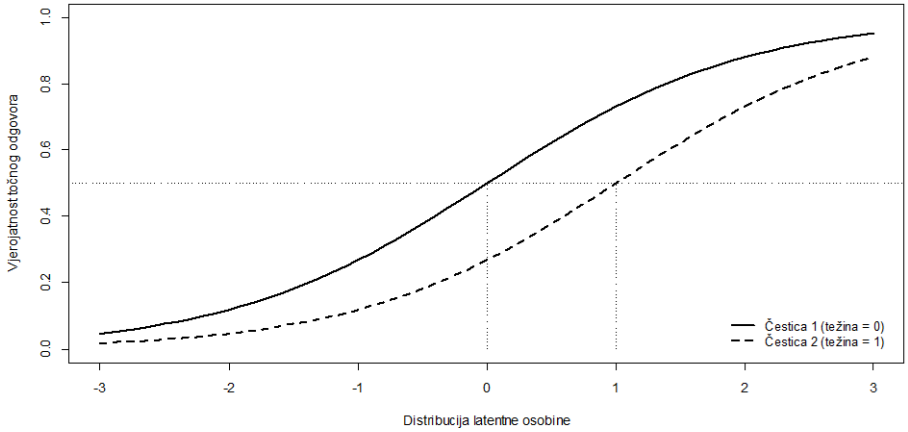
Težina čestice (b_i)

U okviru i CTT-a i IRT-a težina čestice ne označava subjektivnu procjenu toga koliko je zadatak „težak", već opisuje vjerojatnost točnoga odgovora. U CTT-u težina se procjenjuje kao omjer točnih odgovora (za dihotomne čestice) ili prosječna odabrana kategorija (za politomne čestice), zbog čega je izravno ovisna o distribuciji sposobnosti ispitanika. Stoga ista čestica može imati različitu težinu ako na nju odgovaraju adolescenti, starije osobe ili osobe s kliničkim obilježjima, čime se otežavaju usporedba među populacijama i konstrukcija instrumenata.

Nasuprot tomu, IRT uvodi parametar b_i (parametar lokacije ili limen) koji predstavlja onu razinu latentne osobine (θ) pri kojoj je vjerojatnost točnoga odgovora .50. Na Slici 2. prikazane su dvije čestice različite težine: povlačenjem vodoravne crte kroz točku p = .50 do sjecišta s ICC-om te okomite crte na os θ dobiva se vrijednost b_i. Čestica s višom vrijednošću b_i „teža" je jer za svaku razinu θ pruža nižu vjerojatnost točnoga odgovora.

Slika 2.

Prikaz čestica različite težine



Za razliku od CTT-a, parametar težine u IRT-u procjenjuje se neovisno o uzorku jer model opisuje funkcionalni odnos odgovora s obzirom na latentnu osobinu, a ne samo postotak točnih odgovora u određenoj skupini. To omogućuje da se osobine ispitanika i čestica izražavaju na istoj ljestvici, čime se postiže tzv. invarijantnost parametara, pod uvjetom da je model ispravno specificiran. Time rezultati postaju usporedivi među različitim verzijama testa i populacijama, što CTT ne omogućuje (Crocker i Algina, 2008; Hambleton, 1989; Hambleton i Jones, 1993). Ta je invarijantnost parametara i procjena latentnih osobina temelj mnogobrojnih primjena: izjednačavanja i povezivanja rezultata različitih testnih formi, konstrukcije paralelnih verzija s jednakim informacijskim svojstvima, računalno adaptivnoga testiranja (CAT) te identifikacije diferencijalnoga funkcioniranja čestica (DIF).

Međutim, važno je razlikovati tu invarijantnost parametara od strožega pojma specifične objektivnosti, koji vrijedi isključivo u Raschevu (1PL) modelu. Dok invarijantnost znači da su razlike između težina čestica ili sposobnosti ispitanika približno stabilne među različitim uzorcima i testovima, specifična objektivnost podrazumijeva da se usporedbe među osobama mogu raditi potpuno neovisno o odabranim česticama i obrnuto, bez ikakve ovisnosti o drugoj skupini parametara (Baker, 2001; Presser i sur., 2004).

Diskriminativnost čestice (a_i)

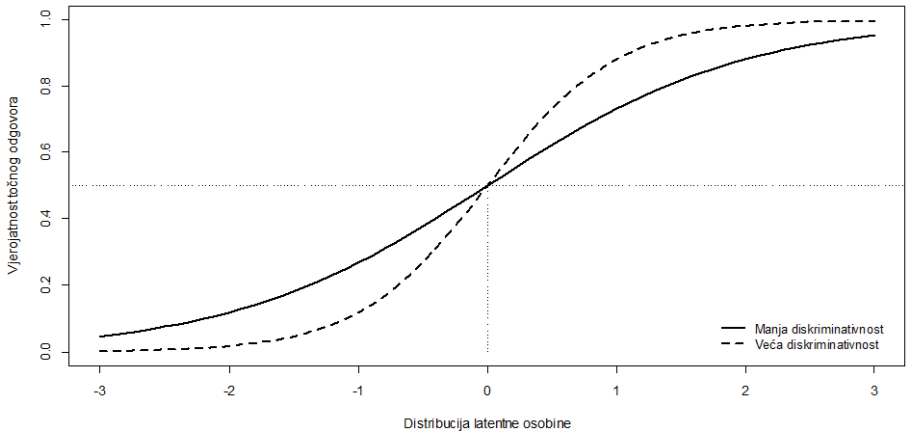
Drugi je ključni parametar u IRT-u diskriminativnost čestice, označena s a_i. Taj parametar opisuje koliko učinkovito čestica diferencira ispitanike koji se razlikuju u razini latentne osobine θ. Što je a_i veći, to je karakteristična krivulja strmija oko točke b_i, što znači da i male promjene u θ dovode do znatnih promjena u vjerojatnosti odgovora.

U praksi se a_i najčešće kreće između 0.5 i 2.5 (Baker, 2001) premda u specifičnim kontekstima, poput kliničkih procjena, mogu biti i ekstremnije vrijednosti (Reise i Waller, 2009). Visoko diskriminativne čestice posebno su korisne u selekciji, dijagnostici i CAT-u, gdje je cilj precizno razlikovati ispitanike u uskome području skale (Chang i Ying, 1999). Međutim, one time mogu smanjiti informativnost izvan toga raspona i otežati generalizaciju na šire populacije. Vrlo visoke vrijednosti a_i mogu upućivati na čestice koje mjere usko definirana iskustva ili specifične stilove odgovaranja, a ne nužno latentni konstrukt (Embretson i Reise, 2000). Kod višedimenzionalnih instrumenata, visoka diskriminativnost može čak prikriti narušenu lokalnu nezavisnost ili prisutnost sekundarnih dimenzija (Embretson i Reise, 2000; Yen, 1993). Stoga a_i treba interpretirati uvijek u kontekstu strukture testa i kvalitete modela (de Ayala, 2009).

Na Slici 3. prikazana je usporedba dviju čestica s različitim razinama a_i: strmija krivulja bolje razlikuje ispitanike kod b_i, dok ona blažega nagiba daje manje informacija u tome području.

Slika 3.

Prikaz čestica različite diskriminativnosti



Parametri procijenjeni pomoću IRT-a i CTT-a uspoređivani su u mnogim istraživanjima. Simulacijske studije Ljubotine (2000) pokazale su da IRT, osobito kada se koriste težinski uravnotežene i visoko diskriminativne čestice, omogućuje precizniju procjenu i manju uvjetnu pogrešku mjerenja u odnosu na CTT. Težine dobivene CTT-om pritom su znatno osjetljivije na sastav uzorka.

Ipak, iako simulacije i teorijske analize (Ljubotina, 2000) jasno ukazuju na prednosti IRT-a u preciznosti, empirijska istraživanja (Courville, 2004; Fan, 1998; Lawson, 1991; Progar i Sočan, 2008) često pokazuju da su procijenjeni parametri u mnogim praktičnim primjenama vrlo slični, osobito kod jednostavnijih IRT modela (1PL, 2PL). Parametri diskriminativnosti pritom pokazuju veću varijabilnost, ali bez

konzistentnih dokaza da IRT pruža stabilnije parametre ili značajno višu valjanost u usporedbi s CTT-om (Fan, 1998; MacDonald i Paunonen, 2002). Stoga primjena IRT-a treba biti uvjetovana ciljevima istraživanja, veličinom uzorka i složenošću instrumenta jer CTT u mnogim slučajevima ostaje praktična i dovoljno pouzdana alternativa.

Ostali parametri čestica

Osim osnovnih parametara težine (b_i) i diskriminativnosti (a_i), prošireni IRT modeli uključuju i dodatne parametre koji omogućuju fleksibilnije modeliranje odgovora ispitanika.

Tako troparametarski model (3PL) uvodi parametar pseudopogađanja (c_i) koji označava donju asimptotu karakteristične krivulje čestice i predstavlja minimalnu vjerojatnost točnoga odgovora čak i kod vrlo niskih razina latentne osobine. Taj je parametar posebno važan u testovima s distraktorima gdje postoji šansa da ispitanik točno odgovori slučajnim pogađanjem (Lord, 1980).

Četveroparametarski model (4PL) dodatno uključuje parametar gornje asimptote (d_i) koji definira maksimalnu vjerojatnost točnoga odgovora, tj. razinu iznad koje vjerojatnost više ne raste, čak ni kod vrlo visokih vrijednosti latentne osobine. Za razliku od klasičnih modela koji pretpostavljaju da ta vrijednost iznosi 1, u nekim kontekstima to nije realistično. Primjerice, čak i vrlo sposobni ispitanici mogu pogriješiti zbog zbunjujućih distraktora, vremenskoga pritiska ili nejasno sročenih zadataka pa se u takvim slučajevima d_i modelira manjim od 1 (Barton i Lord, 1981).

Informacijska krivulja testa

Jedna je od važnijih prednosti IRT-a u odnosu na CTT mogućnost da se preciznost mjerenja analizira u funkciji razine latentne osobine. Preciznost mjerenja tako se može kvantificirati pomoću tzv. informacijske funkcije, i to na razini pojedine čestice (engl. *item information function*, *IIF*) i na razini cijeloga testa (engl. *test information function*, *TIF*). Informacijska funkcija čestice opisuje koliko informacija o ispitanikovoj latentnoj osobini pruža određena čestica na različitim razinama te osobine, dok informacijska funkcija testa zbraja sve pojedine informacijske funkcije čestica i tako kvantificira preciznost mjerenja cjelokupnoga mjernog instrumenta. Grafički prikaz informacijske funkcije naziva se informacijskom krivuljom.

Opća formula za informacijsku funkciju dihotomne čestice (Hambleton i sur., 1991), neovisno o korištenome modelu, glasi:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)\big(1 - P_i(\theta)\big)} \ ,$$

pri čemu je:

- $I_i(\theta)$ – informacija koju čestica *i* pruža za osobu s razinom latentne osobine θ,
- $P_i(\theta)$ – vjerojatnost točnoga odgovora pri razini latentne osobine θ,
- $1 - P_i(\theta)$ – vjerojatnost netočnoga odgovora pri razini latentne osobine θ.

Informacijska funkcija testa dobiva se zbrajanjem informacija svih čestica:

$$I_{\text{test}}(\theta) = \sum_{i=1}^{n} I_i(\theta).$$

Informacijska funkcija u IRT-u analogna je pouzdanosti u CTT-u, no, u odnosu na nju, ima neke bitne prednosti. Naime, pouzdanost se izražava kao jedinstveni koeficijent koji opisuje preciznost rezultata u cijelome rasponu mogućih vrijednosti, odnosno udio varijance pravih rezultata u ukupno izmjerenoj varijanci. Nasuprot tomu, informacijska funkcija preciznost kvantificira lokalno, tj. za svaku vrijednost latentne osobine posebno. Takva razina specifičnosti omogućuje precizno planiranje testnih formi i njihovu optimizaciju čak i prije njihove primjene jer je moguće unaprijed predvidjeti na kojim razinama latentne osobine test pruža najviše informacija. Također omogućuje brzu procjenu učinka uklanjanja ili zamjene pojedinih čestica na ukupnu informativnost instrumenta, kao i procjenu pouzdanosti za definiranu populaciju korištenjem poznatih parametara čestica i pretpostavljene distribucije latentne osobine, bez potrebe za dodatnim prikupljanjem podataka (Embretson i Reise, 2000).

Dok CTT zahtijeva empirijsku primjenu mjernoga instrumenta radi procjene pouzdanosti, IRT omogućuje prediktivnu analizu mjernih karakteristika instrumenata na temelju poznatih parametara čestica. Ta mogućnost sačinjava temelj za razvoj učinkovitijih računalno adaptivnih testova, kao i kraćih verzija postojećih instrumenata (Embretson i Reise, 2000; Lord, 1980; Weiss, 1982).

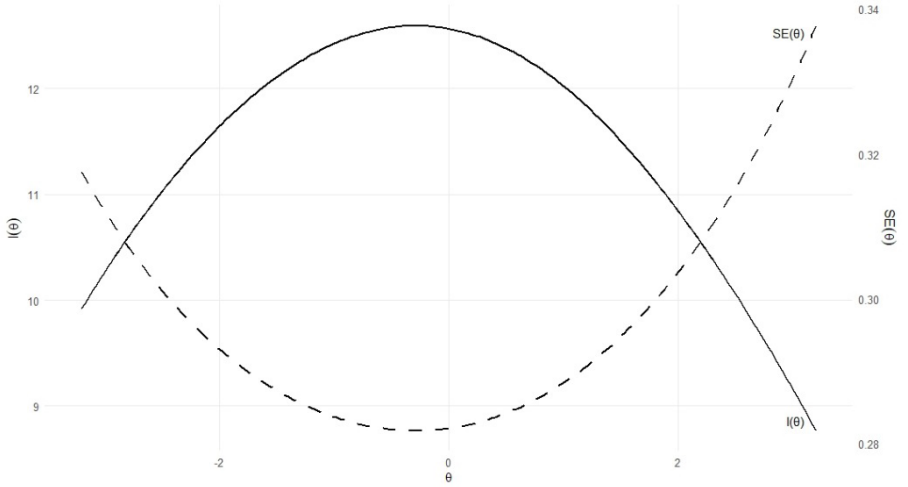
Recipročna vrijednost informacijske funkcije daje standardnu pogrešku mjerenja na toj razini θ:

$$SE(\theta) = \frac{1}{\sqrt{I_{\text{test}}(\theta)}}.$$

Na Slici 4. prikazane su informacijska funkcija testa (puna linija) i standardna pogreška mjerenja (isprekidana linija).

Slika 4.

Informacijska funkcija



Napomena. Informativnost testa I(θ) prikazana je punom linijom, a standardna pogreška mjerenja *SE*(θ) isprekidanom linijom.

Pogreška mjerenja u okviru IRT-a također se može modelirati na svakoj razini latentne osobine zasebno. To je osobito bitno u kliničkoj psihologiji gdje je precizno mjerenje u ekstremnim vrijednostima presudno važno. IRT modelom moguće je tako utvrditi da je određeni upitnik precizniji u mjerenju srednjih razina latentne osobine nego njezinih vrlo niskih ili vrlo visokih razina.

Primjerice, Gray-Little i suradnici (1997) IRT analizom Rosenbergove skale samopoštovanja, inače poznate po visokoj unutarnjoj konzistentnosti, pokazali su da je njezina informativnost vrlo niska za osobe s visokom razinom samopoštovanja, upravo ondje gdje bi precizno razlikovanje bilo ključno u istraživanjima promjena ili usporedbi unutar visoko funkcionalne populacije.

Slično tomu, Reise i Henson (2000) otkrili su da jedna čestica iz Skale anksioznosti Revidiranoga NEO inventara ličnosti pruža gotovo četiri puta više informacija od bilo koje druge čestice u skali, što ukazuje na značajne razlike u informativnosti pojedinih čestica i na važnost njihove pažljive selekcije. Korištenje informacijskih krivulja i uvjetnih pogreški mjerenja tako omogućuje optimizaciju instrumenata za ciljanu populaciju i svrhu mjerenja.

Karakteristična krivulja testa

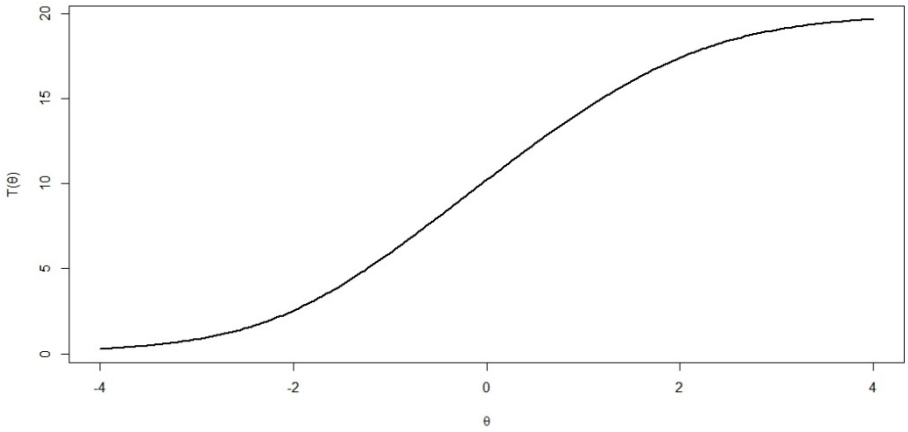
Karakteristična krivulja testa (engl. *test characteristic curve*, TCC) predstavlja zbroj karakterističnih krivulja pojedinačnih čestica i prikazuje očekivani ukupni rezultat testa kao funkciju razine latentne osobine θ. Drugim riječima, ta krivulja

prikazuje kako se prosječan broj bodova koji će ispitanik ostvariti na testu mijenja u odnosu na njegovu latentnu osobinu.

Na Slici 5. prikazan je primjer karakteristične krivulje testa u kojemu se vidi kako porastom razine latentne osobine θ raste i očekivani broj točnih odgovora.

Slika 5.

Karakteristična krivulja testa



Napomena. T(θ) – očekivani ukupni rezultat.

TCC ima sigmoidni (S-oblik), poput karakteristične krivulje pojedinačne čestice, no, za razliku od nje, TCC prikazuje očekivani ukupni rezultat, a ne vjerojatnost točnoga odgovora na jednoj čestici. Očekivani rezultat dobiva se zbrajanjem očekivanih vrijednosti odgovora na svaku česticu (temeljenih na njihovim ICC-ima) za svaku vrijednost latentne osobine θ.

Ta je funkcija osobito korisna u kontekstu interpretacije ukupnih rezultata u metrici latentne osobine, kao i za usporedbu različitih testova koji mjere isti konstrukt, konstruiranje ekvivalentnih formi ili povezivanje rezultata iz različitih verzija mjernih instrumenata (engl. *test linking and equating*; Kolen i Brennan, 1995).

Pretpostavke teorije odgovora na zadatak

Teorija odgovora na zadatak temelji se na nekoliko ključnih pretpostavki koje osiguravaju valjanost procjena latentnih osobina i parametara čestica. Najvažnije među njima su jednodimenzionalnost, lokalna nezavisnost i odgovarajuća specifikacija modela.

Jednodimenzionalnost je osnovna pretpostavka IRT-a koja zahtijeva da sve čestice unutar mjernoga instrumenta mjere jednu zajedničku latentnu osobinu. Ta pretpostavka podrazumijeva da svi ispitanici imaju jednu vrijednost latentne

varijable i da su sve individualne razlike u odgovorima objašnjive isključivo varijacijama u toj jednoj dimenziji. Drugim riječima, model pretpostavlja da je θ jedini sistematski faktor koji utječe na odgovore ispitanika, dok se svi ostali utjecaji (npr. kontekstualni faktori, slučajne pogreške) smatraju slučajnima i nesistematskima (deMars, 2010).

U praksi mnogi instrumenti mjere više od jedne latentne osobine. Ipak, mjerni instrument može biti tehnički jednodimenzionalan čak i ako uključuje više kognitivnih ili psiholoških procesa, pod uvjetom da su ti procesi relativno stabilno integrirani u svim česticama i da ne postoji sustavna varijacija među njima. Primjerice, matematički zadaci riječima mogu ostati jednodimenzionalni ako sve čestice podjednako zahtijevaju i čitalačku sposobnost i matematičko razumijevanje. S druge strane, u slučaju kada postoji namjerna višedimenzionalnost mjernoga instrumenta, preporuča se upotreba višedimenzionalnih IRT modela (Reckase, 2009) te tada, naravno, pretpostavka jednodimenzionalnosti nije uvjet.

Simulacijsko istraživanje koje su proveli Crişan i suradnici (2017) pokazalo je da povrede pretpostavki, poput korištenja jednodimenzionalnog modela za dvodimenzionalno generirane podatke, mogu dovesti do znatnih pristranosti u procjeni latentnih osobina. Te su pristranosti bile posebno izražene na individualnoj razini, što je klinički i praktično važno u dijagnostičkome i selekcijskom kontekstu. Iako su grupne procjene relativno otpornije, preciznost individualnih procjena značajno je opadala, osobito kada su podaci odstupali od pretpostavki modela (Crişan i sur., 2017; Sinharay i Monroe, 2025).

Druga je temeljna pretpostavka IRT-a lokalna nezavisnost. Ona podrazumijeva da su odgovori na pojedine čestice međusobno nezavisni kada se kontrolira razina latentne osobine. To znači da uz istu razinu θ odgovor na jednu česticu ne daje nikakvu dodatnu informaciju o odgovoru na drugu česticu. Ako je ta pretpostavka zadovoljena, sve su korelacije među česticama u potpunosti objašnjene zajedničkom latentnom varijablom. Kršenje pretpostavke lokalne nezavisnosti može ukazivati na prisutnost dodatnih dimenzija ili na povezanost čestica zbog sličnoga sadržaja, položaja u testu, zajedničkih podražaja ili konteksta. Primjerice, čestice koje se odnose na slične situacije ili koje dijele isti tekstualni uvod mogu pokazati rezidualne korelacije iako se modelira samo jedna latentna osobina (Yen, 1993). Za procjenu lokalne nezavisnosti često se koristi Yenov indeks Q_3 (Yen, 1984) koji izračunava korelacije među rezidualima čestica nakon uklanjanja utjecaja latentne varijable.

Treća je pretpostavka IRT-a da je odabrani model odgovarajuće specificiran, tj. da pravilno opisuje odnose između latentne osobine i odgovora na čestice. Procjena dobroga slaganja modela s podacima (engl. *model fit*) uključuje analize na razini čestica (engl. *item fit*), razini pojedinaca (engl. *person fit*) i globalno (engl. *overall fit*). Najčešće korišteni statistički testovi uključuju Pearsonov χ^2, pokazatelj G^2 te njihove modifikacije kao što su S-χ^2 i S-G^2 (deMars, 2010).

Osim kvantitativnih pokazatelja, preporučuje se i vizualna inspekcija (npr. grafički prikazi reziduala ili karakterističnih krivulja čestica) jer može pomoći u

identifikaciji odstupanja koja statistički testovi ne otkrivaju dovoljno jasno (Drasgow i sur., 1995).

U praksi psihometrijski instrumenti rijetko u potpunosti zadovoljavaju stroge pretpostavke IRT-a, zbog čega evaluaciju modela treba temeljiti na stupnju odstupanja, a ne na binarnoj procjeni ispunjava li model sve kriterije ili ne (McDonald, 1981). Važno je pritom razlikovati statističku od praktične značajnosti odstupanja. Iako modeli često ne odgovaraju savršeno podacima, istraživanja pokazuju da su praktične posljedice toga neslaganja često zanemarive; primjerice, razlike u procjeni latentne osobine ili klasifikaciji ispitanika mogu biti minimalne (Sinharay i Haberman, 2014; Zhao i Hambleton, 2017). Međutim, kada su odstupanja od pretpostavki veća i kada model pokazuje slabije slaganje sa stvarnim podacima u smislu i statistički i praktično značajnoga odstupanja, preporučuje se korištenje prikladnijega modela ili uklanjanje čestica koje ne zadovoljavaju pretpostavke (Crişan i sur., 2017; Sinharay i Haberman, 2014). Osim toga, stabilnost modela treba dodatno provjeriti na različitim uzorcima jer dobro slaganje modela u jednoj skupini ne jamči njegovu podjednaku prikladnost u drugoj (Crişan i sur., 2017).

Jednodimenzionalni IRT modeli za dihotomne i politomne čestice

Modeli za dihotomne čestice

Modeli za dihotomne čestice opisuju odnos između latentne osobine i vjerojatnosti točnoga odgovora na česticu s dvjema mogućim kategorijama odgovora (npr. točno – netočno). Razlikuju se prema broju parametara koje uključuju te obliku funkcije kojom se opisuje odnos latentne varijable i vjerojatnosti odgovora (logistička ili funkcija normalne ogive). U literaturi su uobičajene kratice 1PL, 2PL, 3PL i 4PL za jedno- do četveroparametarske logističke modele te 1PNOM, 2PNOM, 3PNOM i 4PNOM za njihove ekvivalente temeljene na normalnoj ogivi.

Jednoparametarski model za predviđanje vjerojatnosti točnoga odgovora koristi isključivo parametar težine čestice (b_i). Dvoparametarski model uz parametar težine uključuje i parametar diskriminativnosti čestice (a_i). Troparametarski model dodatno uvodi parametar donje asimptote (c_i), poznat i kao parametar pseudopogađanja, dok četveroparametarski model, koji se rjeđe koristi, uključuje još i parametar gornje asimptote (d_i).

Iako su modeli s manje parametara ranije razvijeni, danas se na njih gleda kao na posebne slučajeve, odnosno ograničene modele četveroparametarskoga modela u kojima su jedan parametar ili više njih fiksirani na određenu vrijednost (0 ili 1). Zbog toga se teorijska prezentacija često započinje najopćenitijim, četveroparametarskim modelom, iz kojega se svi ostali mogu izvesti ograničavanjem parametara (deMars, 2010).

Matematički je izraz za model 4PL sljedeći:

$$P(\theta) = c_i + (d_i - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}.$$

Model 3PL od prethodnoga se razlikuje samo po tome što se parametar d_i fiksira na vrijednost 1 pa je matematički izraz sljedeći:

$$P(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}.$$

Kod modela 2PL, osim parametra gornje asimptote, fiksira se i parametar donje asimptote na vrijednost 0 pa ta jednadžba izgleda ovako:

$$P(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}.$$

Model 1PL još je jednostavniji te se, osim što su parametri asimptote fiksirani, i variranje parametra diskriminativnosti ograničava tako da je dopušteno samo da *a* ima istu vrijednost za sve čestice. Matematički se to može prikazati na sljedeći način:

$$P(\theta) = \frac{e^{a(\theta - b_i)}}{1 + e^{a(\theta - b_i)}}.$$

U usporedbi s prethodnom formulom, u gornjoj više nema indeksa *i* nego se samo jedna vrijednost *a* koristi za sve čestice. U praksi se ta vrijednost često fiksira na 1, čime se dobiva Raschev model, no u općenitome modelu 1PL dovoljna je pretpostavka da je *a* isti za sve čestice.

Iako se Raschev model (Rasch, 1960/1980) i model 1PL (Birnbaum, 1968) često izjednačavaju zbog matematičke ekvivalentnosti, razlikuju se u pristupu. Rasch je nezavisno razvio model s ciljem potvrde jednake diskriminativnosti čestica. Iz toga se pristupa razvila posebna psihometrijska tradicija, Rascheva analiza, koja inzistira na prilagodbi instrumenata modelu (Bond i Fox, 2015). Suprotno tomu, zagovornici općega IRT-a smatraju da bi modeli trebali biti prilagođeni empirijskim svojstvima čestica, zbog čega je njihov pristup ponajprije eksplorativni (Thissen i Orlando, 2001).

Raschev model i model 1PL imaju nekoliko prednosti u odnosu na složenije modele. Njihova je glavna prednost konceptualna jasnoća interpretacije: za procjenu latentne osobine ispitanika dovoljan je samo broj točnih odgovora, bez potrebe za analiziranjem obrasca odgovaranja. Svi ispitanici s istim ukupnim rezultatom imaju istu procjenu sposobnosti. Kod složenijih modela isti broj točnih odgovora može rezultirati različitim procjenama zbog različitih parametara diskriminativnosti ili pseudopogađanja, što omogućava preciznije mjerenje, ali otežava interpretaciju. Osim toga, ICC krivulje u Raschevu modelu i modelu 1PL paralelne su (u smislu da imaju isti nagib i ne sijeku se iako je riječ o zakrivljenim funkcijama), dok se u modelima 2PL i 3PL mogu križati, što znači da relativna težina čestica ovisi o razini

latentne osobine ispitanika i dodatno komplicira interpretaciju rezultata (deMars, 2010).

Modeli za politomne čestice

Politomne čestice obuhvaćaju one s više od dviju kategorija odgovora, pri čemu kategorije mogu biti ordinalne, poput Likertovih skala, ili nominalne, primjerice, kod pitanja o preferencijama. Za razliku od modela za dihotomne čestice, politomni se modeli ne klasificiraju prema broju parametara već prema prirodi čestica i formatu odgovora koji modeliraju. Njihova je osnovna svrha precizno opisivanje odnosa između latentne osobine (θ) i vjerojatnosti odabira pojedine kategorije. Taj se odnos prikazuje krivuljama kategorija odgovora (engl. *category response curves*, CRC), analognima ICC krivuljama dihotomnih modela. Pritom u stručnoj literaturi i dalje postoji znatna terminološka neujednačenost (de Ayala, 1993, 2009; de Ayala i Sava-Bolesta, 1999; deMars, 2010; Embretson i Reise, 2000; Oshima i Morris, 2008; Sijtsma i Meijer, 2007; Wilson, 2005) koja zahtijeva pažljiv odabir i interpretaciju pojmova u kontekstu korištenoga modela.

Svi su politomni modeli u osnovi ekstenzija modela 2PL za dihotomne čestice, uz različita ograničenja i specifičnosti (van der Linden i Hambleton, 1997).

Najčešće korišten model za ordinalne politomne čestice model je stupnjevitih odgovora (engl. *graded response model*, GRM; Samejima, 1969) koji modelira kumulativne vjerojatnosti odabira određene kategorije odgovora ili kategorija viših od te, a zatim iz tih kumulativnih funkcija posredno izračunava vjerojatnost odabira točno određene kategorije. Ta struktura omogućuje fleksibilno modeliranje ordinalnih skala kod kojih kategorije nisu nužno ravnomjerno raspoređene, no zahtijeva da kumulativne funkcije budu monotono rastuće i paralelne da bi se očuvao smisao ordinalnosti. Njegova pojednostavnjena varijanta, modificirani model stupnjevitih odgovora (Muraki, 1990), uvodi dodatno ograničenje jednakih razmaka među kategorijama za sve čestice, što povećava usporedivost, ali istodobno pretpostavlja da ispitanici uniformno upotrebljavaju skalu, što nije uvijek empirijski opravdano.

Generalizirani model djelomičnih odgovora (engl. *generalized partial credit model*, GPCM; Muraki, 1992) razvijen je s ciljem izravnoga modeliranja vjerojatnosti odabira svake kategorije na temelju odnosa susjednih kategorija. Taj model, za razliku od GRM-a, ne zahtijeva strogu ordinalnost i time pruža veću fleksibilnost, što može biti prednost u instrumentima sa složenijim obrascima odgovora, ali i otežava interpretaciju dobivenih parametara. PCM (engl. *partial credit model*; Masters, 1982), nastao unutar Rascheve tradicije, jednostavnija je varijanta GPCM-a jer pretpostavlja jednaku diskriminativnost svih čestica, čime se smanjuje broj parametara i osigurava parsimonija, no uz cijenu veće restriktivnosti modela.

Još je rigorozniji u tome pogledu model skala ocjenjivanja (engl. *rating scale model*, RSM; Andrich, 1978) koji, uz pretpostavku fiksne diskriminativnosti, predviđa i identične limene te jednake razmake među kategorijama za sve čestice. Taj je pristup osobito primjeren kada sve čestice imaju istovjetan format, primjerice, u klasičnim Likertovim ljestvicama, i kada se očekuje da ispitanici dosljedno koriste ponuđene kategorije. Međutim, u situacijama kada se čestice razlikuju po težini ili u obrascu korištenja kategorija primjena RSM-a može biti neadekvatna jer model nije dovoljno fleksibilan da te razlike zahvati.

S druge strane, model nominalnih odgovora (engl. *nominal response model*, NRM; Bock, 1972) konstruiran je upravo za slučajeve u kojima kategorije odgovora nemaju prirodni redoslijed. Svaka kategorija u tome modelu ima vlastiti skup parametara koji određuju njezin položaj i oblik odgovora u odnosu na latentnu osobinu. NRM je izvorno razvijen da bi omogućio psihometrijsku analizu distraktora u zadacima višestrukoga izbora, čime pogrešni odgovori postaju relevantni pokazatelji ispitanikova ponašanja, a ne tek „šum". Danas se primjenjuje ne samo za nominalne čestice, već i za empirijsku provjeru ordinalnosti u politomnim instrumentima te za modeliranje testleta kao jedinstvenih jedinica odgovora (Samejima, 1988, 1996; Wainer i Kiely, 1987). Zbog svoje je fleksibilnosti koja omogućuje procjenu diskriminativnosti i lokacije svake pojedine kategorije bez pretpostavki o međusobnome redoslijedu posebno vrijedan u kontekstu dijagnostike, ispitivanja preferencija ili zadataka kategorizacije.

Važno je naglasiti da se parametri procijenjeni pomoću različitih politomnih modela ne mogu izravno međusobno uspoređivati jer svaki od njih polazi od specifičnih pretpostavki o strukturi kategorija i načinu na koji latentna osobina utječe na izbor odgovora. Izbor odgovarajućega modela primarno ovisi o prirodi podataka i strukturi instrumenta. Kod podataka bez jasne ordinalnosti najprikladnije je započeti s nominalnim modelom (Foster i sur., 2017), dok u slučajevima varijabilne diskriminativnosti čestica GRM i Murakijevi modeli (1990, 1992) nude veću fleksibilnost u odnosu na Rascheve pristupe. RSM se preporučuje samo kada sve čestice koriste identičnu ljestvicu bodovanja (deMars, 2010).

Iako teorijski kriteriji pružaju važne smjernice za odabir modela, konačnu odluku treba potkrijepiti empirijskom provjerom, primjerice, usporedbom pokazatelja slaganja podataka s modelom poput log-vjerojatnosti ili pokazatelja χ^2. U praksi odabir često ovisi i o pragmatičnim razlozima kao što su veličina uzorka, jednostavnost bodovanja, dostupnost softvera i filozofija mjerenja (Baker, 2001; Embretson i Reise, 2000).

Primjena IRT-a u psihologiji

Teorija odgovora na zadatak prvi je put šire primijenjena u velikim obrazovnim testiranjima, gdje je ponudila rješenja za neka ključna ograničenja CTT-a. To se

osobito odnosi na problem usporedivosti rezultata ispitanika koji su rješavali različite verzije istoga testa. Primjena IRT-a u standardiziranim ispitivanjima za upis na američke fakultete (npr. SAT, GRE) u američkome nacionalnome obrazovnom testiranju (NAEP; Oranje i Kolstad, 2019; von Davier i sur., 2005) te u međunarodnim istraživanjima poput PISA-e i TIMSS-a (OECD, 2017, 2024) omogućila je kalibraciju čestica i testova na zajedničkoj ljestvici te usporedbu ispitanika neovisno o korištenoj verziji ili populaciji kojoj pripadaju. Ključnu ulogu u tome procesu imaju tzv. vezne čestice (engl. *linking items*) koje se pojavljuju u više formi testa i služe kao referentna točka za postavljanje rezultata na istu skalu.

IRT nije samo unaprijedila preciznost mjerenja, nego je omogućila i razvoj kraćih, učinkovitijih i pravednijih testova. Jedno od važnih područja u kojima se to vidi analiza je diferencijalnoga funkcioniranja čestice (DIF) koja je IRT pristup provjeri mjerne invarijantnosti među skupinama. DIF analize otkrivaju čestice na koje ispitanici iz različitih skupina, unatoč jednakoj razini latentne osobine, odgovaraju s različitom vjerojatnošću, tj. one čestice koje potencijalno favoriziraju ili diskriminiraju određene skupine, čime se doprinosi razvoju kulturno i demografski osjetljivijih instrumenata (Millsap, 2011; Zumbo, 1999).

IRT također sačinjava temelj za računalno adaptivno testiranje koje bitno povećava učinkovitost i preciznost mjerenja (Embretson i Reise, 2000; Zhang i Chang, 2016). U CAT-u se izbor svake nove čestice prilagođava prethodnim odgovorima ispitanika, a test se nastavlja dok se ne postigne zadana razina preciznosti. Time se testovi skraćuju, povećava se njihova efikasnost, a smanjuje se mogućnost varanja radi prepoznavanja zadataka, što je važno u uvjetima *online* testiranja (Hambleton, 1991; Weiss, 1982, 2004).

Kao nadogradnja CAT-a sve se češće koristi višestupanjsko testiranje (engl. *multistage testing*, MST) koje omogućuje bolju kontrolu sadržaja. Dok CAT bira svaku česticu pojedinačno, MST strukturira test u module različitih težina. Nakon inicijalnoga modula ispitanici prelaze na lakši ili teži modul, ovisno o postignutome rezultatu, čime se kombiniraju adaptivnost i veća sadržajna kontrola (Yan i sur., 2014; Zhang i Chang, 2016).

Osim u „pametnome testiranju", IRT se sve češće primjenjuje i u tzv. „pametnome učenju" (engl. *smart learning*). Tu se IRT koristi za praćenje znanja učenika, prilagodbu nastavnoga sadržaja njihovoj razini sposobnosti te otkrivanje područja koja zahtijevaju dodatnu podršku. Na temelju obrazaca odgovora moguće je automatizirano preporučiti zadatke odgovarajuće težine i kreirati personalizirane obrazovne putove, čime se povećava angažman i učinkovitost u učenju. Platforme kao što su *Duolingo* i *Khan Academy* koriste upravo IRT modele za dinamičko prilagođavanje sadržaja znanju i vještinama korisnika (Zhang i Chang, 2016).

IRT se sve češće primjenjuje i u drugim područjima psihologije, poput kliničke dijagnostike, mjerenja osobina ličnosti i evaluacije zdravstvenih ishoda (Cella i sur., 2010; Reise i Waller, 2009). U tome se području ističe projekt PROMIS (*Patient Reported Outcomes Measurement Information System*) koji su razvili američki

Nacionalni zdravstveni instituti (National Institutes of Health, 2012). PROMIS omogućuje stručnjacima u području zdravlja praćenje učinkovitosti intervencija i tretmana pomoću standardiziranih skala za tjelesno i mentalno zdravlje i upotrebe CAT-a. Jedno je od glavnih postignuća toga projekta objedinjavanje čestica iz različitih instrumenata u zajedničku bazu (tzv. *item banking*) te njihova kalibracija na jedinstvenoj skali, čime se omogućuje usporedivost rezultata među istraživanjima koja koriste različite mjere i populacije.

Psihometrijska svojstva mnogih kliničkih mjernih instrumenata sve se češće ispituju pomoću IRT-a, pri čemu se najčešće primjenjuje model 2PL. To proizlazi iz činjenice da klinički upitnici rijetko sadrže čestice jednake diskriminativnosti, što ograničava primjenu jednoparametarskih modela. S druge strane, model 3PL koji uključuje parametar pseudopogađanja rijetko se koristi u kliničkome kontekstu zbog nejasne konceptualizacije toga parametra u području mjerenja ličnosti i psihopatologije (Thomas, 2011). Iako su neki autori predlagali da se donja asimptota može interpretirati kao indikator pristranosti u odgovaranju, primjerice, socijalno poželjnoga odgovaranja (Zumbo i sur., 1997), takvo tumačenje ima bitna metodološka ograničenja jer je parametar *c* svojstvo čestice, a ne osobe. Reise i Waller (2003, 2010) tako upozoravaju da se u kontekstu psihopatologije donja i gornja asimptota često pojavljuju zbog toga što neke čestice mjere ekstremne simptome ili iz razloga što čestica ima različito značenje za osobe s kliničkim simptomima i one bez njih. Ekstremni simptomi poput halucinacija možda neće biti prisutni čak ni kod osoba s visokom razinom latentne osobine (primjerice, kod psihoza), dok, s druge strane, simptomi koji su česti u općoj populaciji, poput nesanice ili tjeskobe, mogu biti prisutni i pri niskim razinama psihopatologije. Tako su zapravo karakteristike čestica odgovorne za krivulje s niskim ili visokim asimptotama, a ne prisustvo pristranosti u odgovaranju.

Iako se tek odnedavno primjenjuje u području mjerenja ličnosti, IRT je u tome području omogućio razvoj kraćih i preciznijih verzija postojećih mjernih instrumenata. Primjerice, Maples i suradnici (2014) primijenili su IRT da bi inventar ličnosti IPIP-NEO-300 skratili u verziju sa 120 čestica, pri čemu su pokazali visoku pouzdanost i konvergentnu valjanost u odnosu na originalni instrument NEO-PI-R. Slično tomu, Colledani i suradnici (2018) primijenili su IRT za optimizaciju EPQ-a, čime su također postigli skraćivanje instrumenta uz očuvanje njegovih psihometrijskih svojstava.

U organizacijskoj se psihologiji IRT osobito primjenjuje u selekciji zaposlenika, procjeni kompetencija i evaluaciji radnoga učinka (Craig i Harvey, 2004; Lang i Tay, 2021). Da bi se smanjio utjecaj socijalno poželjnoga odgovaranja, u selekciji se često koriste čestice s prisilnim izborom koje ispitaniku nude tvrdnje jednake poželjnosti, ali različitih latentnih vrijednosti. Kada se takvi testovi boduju prema CTT-u, rezultati postaju ipsativni (samonormitani, to jest usporedivi samo unutar pojedinca) i time ograničavaju međusobne usporedbe (Lin i sur., 2023). Razvojem IRT modela za prisilni izbor omogućena je procjena koja daje rezultate s pravom varijancom,

pogodnom za usporedbu među kandidatima (Brown, 2016; Stark i sur., 2012). Simulacijska istraživanja pokazala su da CAT upitnici s prisilnim izborom mogu postići jednaku točnost procjene latentne varijable s upola manje čestica od neadaptivnih instrumenata (Joo i sur., 2019; Stark i sur., 2012). U praksi se već koriste mnogobrojni takvi mjerni instrumenti, uključujući *Navy Computer Adaptive Personality Scales* (Houston i sur., 2006), *Tailored Adaptive Personality Assessment System* (Drasgow i sur., 2012) i *Adaptive Employee Personality Test* (Boyce i sur., 2014).

Specifičnosti upitnika samoprocjene i mjera tipičnoga ponašanja dovele su do sve češće primjene modela idealne točke (engl. *ideal point* ili *unfolding models*) umjesto uobičajenijih modela dominacije (engl. *dominance models*) razvijenih za mjerenje maksimalnoga učinka (Brown i Maydeu-Olivares, 2010; Stark i sur., 2006). Za razliku od modela dominacije koji pretpostavljaju da se vjerojatnost odabira čestice stalno povećava s rastom latentne osobine, modeli idealne točke polaze od toga da je najveća vjerojatnost odabira kada je položaj čestice na latentnome kontinuumu najbliži stvarnoj razini osobine ispitanika, a zatim opada. Stoga njihova karakteristična krivulja ima zvonolik, a ne sigmoidan oblik (Liu i Wang, 2018). Dok su modeli dominacije prikladni za kognitivne sposobnosti, kod mjerenja tipičnoga ponašanja često nije jasno koji pristup bolje odgovara. Stark i suradnici (2006) pokazali su da se neke čestice iz inventara ličnosti 16PF bolje uklapaju u modele idealne točke, a pogrešan odabir modela može značajno utjecati na rangiranje ispitanika, što ima ozbiljne implikacije za, primjerice, selekciju zaposlenika. Zbog toga se preporučuje da se pri konstrukciji i analizi instrumenata za mjerenje tipičnoga ponašanja empirijski provjeri koji model bolje pristaje podacima da bi se smanjio rizik pogrešne procjene ili nepravednoga rangiranja kandidata.

Kod mjera tipičnoga ponašanja važno je novo područje istraživanja i upotrebe IRT-a i analiza procesa odgovaranja ispitanika, uključujući otkrivanje neautentičnih ponašanja poput upravljanja dojmom ili slučajnoga zaokruživanja (Liu i Zhang, 2020; Stark i sur., 2001; Zickar i Robie, 1999). Tako IRT nadilazi samu procjenu latentnih osobina te postaje i alat za prepoznavanje mentalnih procesa ispitanika pri odgovaranju, što je ključno za osiguravanje kvalitete podataka u istraživačkim i selekcijskim kontekstima.

Unatoč svojim mnogim prednostima primjena IRT-a relativno je ograničena u odnosu na potencijal koji nudi. Jedna je od ključnih prepreka široj prihvaćenosti IRT-a zahtjev za relativno velikim uzorcima ispitanika za pouzdanu procjenu parametara čestica. Tako se u literaturi pojavljuju preporuke o najmanje 250 do 500 ispitanika (npr. deMars, 2010) u jednostavnijim modelima ili pak omjeri ispitanika i parametara u modelu od 10 : 1 ili čak 20 : 1 (de Ayala i Sava-Bolesta, 1999; deMars, 2003).

No, različita simulacijska istraživanja jasno pokazuju da ne postoji jedinstveno pravilo o minimalnoj veličini uzorka potrebnoj za primjenu IRT modela jer ta potreba u velikoj mjeri ovisi o specifičnim obilježjima svakoga istraživanja. Dok su neka istraživanja pokazala da uz korištenje *a priori* distribucija ili metoda procjene koje

su manje osjetljive na prisustvo podataka koji nedostaju stabilne procjene parametara mogu biti moguće i s relativno malim uzorcima (pa i manjima od 100 ispitanika), drugi radovi ukazuju na to da u složenijim modelima čak ni vrlo veliki uzorci od nekoliko tisuća sudionika ne jamče zadovoljavajuću preciznost procjena. Zbog toga se istraživačima preporuča da provedu vlastite simulacije Monte Carlo za procjenu potrebnoga broja ispitanika prije provedbe istraživanja da bi osigurali da su njihovi podaci dostatni za pouzdanu primjenu odabranoga IRT modela i valjanu interpretaciju rezultata (Schroeders i Gnambs, 2025).

Drugo ograničenje odnosi se na složenost analitičkoga postupka, što uključuje potrebu za specijaliziranim statističkim znanjima i softverom. Iako se dostupnost besplatnih alata poput R-a kontinuirano povećava, inicijalna krivulja učenja može biti prilično strma. Nadalje, složeniji modeli zahtijevaju detaljne provjere valjanosti pretpostavki, poput jednodimenzionalnosti ili lokalne nezavisnosti, što dodatno komplicira postupak.

Stoga, unatoč jasno istaknutim prednostima, važno je pažljivo procijeniti uvjete pod kojima primjena IRT-a donosi stvarne koristi u odnosu na jednostavnije, tradicionalne pristupe psihometrijskomu mjerenju.

Zaključak

IRT je suvremeni psihometrijski pristup koji u odnosu na klasične metode nudi mnogobrojne prednosti, uključujući preciznije mjerenje latentnih osobina, detaljno modeliranje karakteristika čestica te zajedničko skaliranje ispitanika i čestica na istoj ljestvici. Zahvaljujući tim svojstvima, IRT omogućuje usporedivost rezultata dobivenih različitim verzijama testova, konstrukciju paralelnih instrumenata s jednakim informacijskim karakteristikama, računalno adaptivno testiranje koje povećava učinkovitost i smanjuje opterećenje za ispitanike te identifikaciju čestica koje ne funkcioniraju jednako u različitim skupinama, čime se osigurava pravednost i valjanost mjerenja.

Ipak, primjena IRT-a suočava se i s izazovima: od potrebe za većim uzorcima i složenijim statističkim znanjem, do metodoloških pitanja povezanih s modeliranjem višedimenzionalnih konstrukata i politomnih čestica. Stoga je ključno primjenu IRT-a prilagoditi ciljevima istraživanja, dostupnim resursima i prirodi podataka da bi njegove prednosti došle do punoga izražaja.

Unatoč tim ograničenjima potencijal IRT-a za psihologiju ostaje iznimno velik. Daljnji razvoj višedimenzionalnih i bayesijskih pristupa te njihova dostupnost kroz otvorene statističke alate dodatno će ojačati poziciju IRT-a kao temelja suvremene psihometrijske prakse, doprinoseći preciznijemu mjerenju i većoj znanstvenoj rigoroznosti unutar discipline.

Literatura

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*, 581–594. https://doi.org/10.1177/014662167800200413

Baker, F. B. (2001). *The basics of item responses theory*. ERIC Clearinghouse on Assessment and Evaluation.

Barton, M. A. i Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model*. Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. U: F. M. Lord i M. R. Novick (Ur.), *Statistical theories of mental test scores* (str. 397–479). Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51. https://doi.org/10.1007/BF02291411

Bond, T. G. i Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge / Taylor & Francis Group.

Boyce, A. S., Conway, J. S. i Caputo, P. M. (2014). *ADEPT-15 technical documentation: Development and validation of Aon Hewitt's personality model and adaptive employee test (ADEPT-15)*. Aon Hewitt.

Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, *81*(1), 135–160. https://doi.org/10.1007/s11336-014-9434-9

Brown, A. i Maydeu-Olivares, A. (2010). Issues that should not be overlooked in the dominance versus ideal point controversy. *Industrial and Organizational Psychology*, *3*(4),489–493. https://doi.org/10.1111/j.1754-9434.2010.01277.x

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., Devellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Lai, J. S., Pilkonis, P., Revicki, D., Rose, M., … PROMIS Cooperative Group (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of clinical epidemiology*, *63*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011

Chang, C.-H. i Reeve, B. B. (2005). Item response theory and its applications to patient-reported outcomes measurement. *Evaluation and the Health Professions*, *28*, 264–282. https://doi.org/10.1177/0163278705278275

Chang, H.-H. i Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211–222. https://doi.org/10.1177/01466219922031338

Colledani, D., Anselmi, P. i Robusto, E. (2018). Using item response theory for the development of a new short form of the Eysenck Personality Questionnaire-Revised. *Frontiers in Psychology*, *9*, članak 1834. https://doi.org/10.3389/fpsyg.2018.01834

Courville, T. G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics*. [Doktorska disertacija]. Texas A & M University, TX, USA. Federated Electronic Theses and Dissertations. http://hdl.handle.net/1969.1/1064

Craig, S. B. i Harvey, R. J. (2004). Using CAT to reduce administration time in 360° performance assessment. U: B. Craig (Chair), *360, The next generation: Innovations in multisource performance assessment*. Symposium presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Chicago.

Crişan, D. R., Tendeiro, J. N. i Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement*, *41*(6), 439–455. https://doi.org/10.1177/0146621617695522

Crocker, L. M. i Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

de Ayala, R. J. (1993). An introduction to polytomous item response theory models. *Measurement and Evaluation in Counseling and Development*, *25*, 172–189.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

de Ayala, R. J. i Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement*, *23*, 3–19. https://doi.org/10.1177/01466219922031130

deMars, C. E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement*, *27*(4), 275–288. https://doi.org/10.1177/0146621603027004003

deMars, C. (2010). *Item response theory*. Oxford University Press.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B. i Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*(2), 143–166. https://doi.org/10.1177/014662169501900203

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L. i White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support army selection and classification decisions (Tech. rep. no. 1311)*. U. S. Army Research Institute for the Behavioral and Social Sciences. https://apps.dtic.mil/sti/tr/pdf/ADA564422.pdf

Embretson, S. E. i Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *3*, 357–381. https://doi.org/10.1177/0013164498058003001

Foster, G. C., Min, H. i Zickar, M. J. (2017). Review of item response theory practices in organizational research: Lessons learned and paths forward. *Organizational Research Methods*, *20*(3), 465–486. https://doi.org/10.1177/1094428116689708

Gray-Little, B., Williams, V. S. L. i Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *23*, 443–451. https://doi.org/10.1177/0146167297235001

Hambleton, R. K. (1989). Principles and selected applications of item response theory. U: R. L. Linn (Ur.), *Educational measurement* (3rd ed., str. 147–200). Macmillan Publishing Co, Inc; American Council on Education.

Hambleton, R. K. (1991). *Fundamentals of item response theory* (Vol. 2). Sage Publications.

Hambleton, R. K. i Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *3*, 38–47. https://doi.org/10.1111/j.1745-3992.199

Hambleton, R. K., Swaminathan, H. i Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications, Inc.

Houston, J. S., Borman, W. C., Farmer, W. L. i Bearden, R. M. (Ur.) (2006). *Development of the Navy Computer Adaptive Personality Scales (NCAPS) (No. NPRST-TR-06–2)*. NavyPersonnel Research, Studies, and Technology Division, Bureau of Naval Personnel (NPRST/PERS-1). https://apps.dtic.mil/sti/tr/pdf/ADA456977.pdf

Joo, S., Lee, P. i Stark, S. (2019). Adaptive testing with the GGUM-RANK multidimensionalforced choice model: Comparison of pair, triplet, and tetrad scoring. *Behavior Research Methods*, *52*, 761–772. https://doi.org/10.3758/s13428-019-01274-6

Kolen, M. J. i Brennan, R. L. (1995). *Test equating: Methods and practices*. SpringerVerlag

Lang, J. W. B. i Tay, L. (2021). The science and practice of item response theory in organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*, 311–338. https://doi.org/10.1146/annurev-orgpsych-012420-061705

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? U: B. Thompson (Ur.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, str. 159–168). JAI.

Lin, Y., Brown, A. i Williams, P. (2023). Multidimensional forced-choice CAT with dominance items: An empirical comparison with optimal static testing under different desirability matching. *Educational and Psychological Measurement*, *83*(2), 322–350. https://doi.org/10.1177/00131644221077637

Liu, C.-W. i Wang, W.-C. (2018). A general unfolding IRT Model for multiple response styles. *Applied Psychological Measurement*, *43*(3), 195–210. https://doi.org/10.1177/0146621618762743

Liu, J. i Zhang, J. (2020). An item-level analysis for detecting faking on personality tests: appropriateness of ideal point item response theory models. *Frontiers in Psychology*, *10*, članak 3090. https://doi.org/10.3389/fpsyg.2019.03090

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Inc.

Ljubotina, D. (2000). *Usporedba psihometrijskih karakteristika kompozitnih testova konstruiranih u sklopu klasične teorije i teorije odgovora na zadatke* [Neobjavljena doktorska disertacija]. Filozofski fakultet, Sveučilište u Zagrebu, Hrvatska.

MacDonald, P. i Paunonen, S. (2002). A Monte Carlo Comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*(6), 921–943. https://doi.org/10.1177/0013164402238082

Maples, J. L., Guan, L., Carter, N. T. i Miller, J. D. (2014). A test of the international personality item pool representation of the revised NEO personality inventory and development of a 120-item IPIP-based measure of the five-factor model. *Psychological Assessment*, *26*(4), 1070–1084. https://doi.org/10.1037/pas0000004

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 100–117. https://10.1111/j.2044-8317.1981.tb00621.x

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge. https://doi.org/10.4324/9780203821961

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*, 59–71. https://doi.org/10.1177/014662169001400106

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176. https://doi.org/10.1177/014662169201600206

Oshima, T. C. i Morris, S. B. (2008). An NCME instructional module on Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice*, *27*(3), 43–50. https://doi.org/10.1111/j.1745-3992.2008.00127.x

National Institutes of Health. (2012). *PROMIS: Patient-Reported Outcomes Measurement Information System*. https://commonfund.nih.gov/promis/index

OECD. (2017). *PISA 2015 Technical Report*. OECD Publishing. https://www.oecd.org/en/about/programmes/pisa/pisa-data.html

OECD. (2024). *PISA 2022 Technical Report*. OECD Publishing. https://doi.org/10.1787/01820d6d-en

Oranje, A. i Kolstad, A. (2019). Research on psychometric modeling, analysis, and reporting of the National Assessment of Educational Progress. *Journal of Educational and Behavioral Statistics*, *44*(6), 671–689. https://doi.org/10.3102/1076998619867105

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M. i Singer, E. (Ur.). (2004). *Methods for testing and evaluating survey questionnaires*. Wiley.

Progar, Š. i Sočan, G. (2008). An empirical comparison of item response theory and classical test theory. *Psihološka Obzorja / Horizons of Psychology*, *17*(3), 5–24.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Expanded edition with foreword and afterword by B. D. Wright, Chicago: The University of Chicago Press). Danish Institute for Educational Research.

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.

Reise, S. P. i Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI–R. *Assessment*, *7*, 347–364. https://doi.org/10.1177/107319110000700404

Reise, S. P. i Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, *8*, 164–184. https://doi.org/10.1037/1082-989X.8.2.164

Reise, S. P. i Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553

Reise, S. P. i Waller, N. G. (2010). Measuring psychopathology with nonstandard item response theory models: Fitting the four-parameter model to the Minnesota Multiphasic Personality Inventory. U: S. E. Embretson (Ur.), *Measuring psychological constructs: Advances in model-based approaches* (str. 147–173). American Psychological Association.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*(S1), 1–97. https://doi.org/10.1007/BF03372160

Samejima, F. (1988). Comprehensive latent trait theory. *Behaviormetrika*, *15*, 1–24. https://doi.org/10.2333/bhmk.15.24_1

Samejima, F. (1996). Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika*, *23*, 17–35. https://doi.org/10.2333/bhmk.23.17

Schroeders, U. i Gnambs, T. (2025). Sample-size planning in item-response theory: A tutorial. *Advances in Methods and Practices in Psychological Science*, *8*(1), 1–13. https://doi.org/10.1177/25152459251314798

Sijtsma, K. i Meijer, R. R. (2007). Nonparametric item response theory and special topics. U: C. R. Rao i S. Sinharary (Ur.), *Handbook of Statistics Vol. 26: Psychometrics* (str. 719–746). Elsevier.

Sinharay, S. i Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*, *33*(1), 23–35. https://doi.org/10.1111/emip.12024

Sinharay, S. i Monroe, S. (2025). Assessment of fit of item response theory models: A critical review of the status quo and some future directions. *British Journal of Mathematical and Statistical Psychology*. Advance online publication. https://doi.org/10.1111/bmsp.12378

Stark, S., Chernyshenko, O. S., Chan, K.-Y., Lee, W. C. i Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology*, *86*(5), 943–953. https://doi.org/10.1037/0021-9010.86.5.943

Stark, S., Chernyshenko, O. S., Drasgow, F. i White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods*, *15*(3), 463–487. https://doi.org/10.1177/1094428112444611

Stark, S., Chernyshenko, O. S., Drasgow, F. i Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal-point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*(1), 25–39. https://doi.org/10.1037/0021-9010.91.1.25

Takšić, V. (2002). Upitnici emocionalne inteligencije (kompetentnosti). U: K. Lacković Grgin, A. Bautović, V. Čubela i Z. Penezić (Ur.), *Zbirka psihologijskih skala i upitnika* (str. 27–45). Filozofski fakultet u Zadru, Hrvatska.

Thissen, D. i Orlando, M. (2001). Item response theory for items scored in two categories. U: D. Thissen i H. Wainer (Ur.), *Test scoring* (str. 73–140). Lawrence Erlbaum Associates Publishers.

Thomas M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment*, *18*(3), 291–307. https://doi.org/10.1177/1073191110374797

van der Linden, W. J. i Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer.

von Davier, M., Sinharay, S., Oranje, A. i Beaton, A. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. U: C. R. Rao i S. Sinharay (Ur.), *Handbook of Statistics: Psychometrics* (Sv. 26, str. 125–194). Elsevier. https://doi.org/10.1016/S0169-7161(06)26032-2

Wainer, H. i Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185–201. http://www.jstor.org/stable/1434630

Ware, J. E. Jr. i Sherbourne, C. D. (1992). The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Medical Care*, *30*(6), 473–483. https://doi.org/10.1097/00005650-199206000-00002

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473–492. https://doi.org/10.1177/014662168200600408

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, *37*(2), 70–84. https://doi.org/10.1080/07481756.2004.11909753

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Yan, D., von Davier, A. A. i Lewis, C. (2014). *Computerized multistage testing: Theory and applications*. CRC Press.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *2*, 125–145. https://doi.org/10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zhang, S. i Chang, H. H. (2016). From smart testing to smart learning: How testing technology can assist the new generation of education. *International Journal of Smart Technology and Learning*, *1*(1), 67–92. https://doi.org/10.1504/IJSMARTTL.2016.078162

Zhao, Y. i Hambleton, R. K. (2017). Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Frontiers in Psychology*, *8*, članak 484. https://doi.org/10.3389/fpsyg.2017.00484

Zickar, M. J. i Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology*, *84*(4), 551–563. https://doi.org/10.1037/0021-9010.84.4.551

Zumbo, B. D. (1999*). A Handbook on the theory and methods of Differential Item Functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense, Ottawa, ON.

Zumbo, B. D., Pope, G. A., Watson, J. E. i Hubley, A. M. (1997). An empirical test of Roskam's conjecture about the interpretation of an ICC parameter in personality inventories. *Educational and Psychological Measurement*, *57*, 963–969. https://doi.org/10.1177/0013164497057006006

What Do Items Really Measure? An Introduction to Item Response Theory

Abstract

Item Response Theory (IRT) is a modern psychometric framework that models the probability of item responses as a function of one or more latent traits. This paper provides a comprehensive overview of the theoretical foundations, assumptions, and applications of IRT, focusing on its use in psychological measurement. The aim of this review is to present key IRT models for both dichotomous and polytomous items, compare them to Classical Test Theory (CTT), and discuss their advantages. The review explains item parameters (difficulty, discrimination, guessing), describes item and test characteristic curves, and discusses key assumptions such as unidimensionality and local independence. Special attention is given to the application of IRT across different areas of psychology and its contributions to psychometric practice, such as ensuring parameter invariance, enhancing measurement precision, and detecting differential item functioning. Challenges related to IRT, such as the need for larger samples, complex analyses, and specific issues with non-cognitive measures, are also addressed. In conclusion, despite its practical demands, IRT remains a fundamental tool for developing contemporary psychological instruments, offering greater accuracy and scientific rigor in psychological assessment.

Keywords: Item Response Theory, psychological measurement, models for dichotomous items, polytomous models, psychometrics

Primljeno: 27. 4. 2025.